Proc. Natl. Acad. Sci. USA
Vol. 96, pp. 2937–2942, March 1999
Genetics

Detecting patterns of protein distribution and gene expression *in silico*

(peroxisomes/peroxisome biogenesis/membrane proteins/lysine synthesis/green fluorescent protein)

MICHAEL T. GERAGHTY*, DOUG BASSETT†‡, JAMES C. MORRELL§, GREGORY J. GATTO, JR.§, JIANWU BAI§, BRIAN V. GEISBRECHT§, PHIL HIETER†¶, AND STEPHEN J. GOULD§‖

Departments of *Pediatrics, †Molecular Biology and Genetics, and §Biological Chemistry, The Johns Hopkins University School of Medicine, Baltimore, MD 21205

Edited by Sherman M. Weissman, Yale University School of Medicine, New Haven, CT, and approved December 30, 1998 (received for review September 9, 1998)

ABSTRACT **Most biological information is contained within gene and genome sequences. However, current methods for analyzing these data are limited primarily to the prediction of coding regions and identification of sequence similarities. We have developed a computer algorithm, CoSMoS (for context sensitive motif searches), which adds context sensitivity to sequence motif searches. CoSMoS was challenged to identify genes encoding peroxisome-associated and oleate-induced genes in the yeast *Saccharomyces cerevisiae*. Specifically, we searched for genes capable of encoding proteins with a type 1 or type 2 peroxisomal targeting signal and for genes containing the oleate-response element, a *cis*-acting element common to fatty acid-regulated genes. CoSMoS successfully identified 7 of 8 known PTS-containing peroxisomal proteins and 13 of 14 known oleate-regulated genes. More importantly, CoSMoS identified an additional 18 candidate peroxisomal proteins and 300 candidate oleate-regulated genes. Preliminary localization studies suggest that these include at least 10 previously unknown peroxisomal proteins. Phenotypic studies of selected gene disruption mutants suggests that several of these new peroxisomal proteins play roles in growth on fatty acids, one is involved in peroxisome biogenesis and at least two are required for synthesis of lysine, a heretofore unrecognized role for peroxisomes. These results expand our understanding of peroxisome content and function, demonstrate the utility of CoSMoS for context-sensitive motif scanning, and point to the benefits of improved *in silico* genome analysis.**

The advent of expressed sequence tag and genome sequencing projects has greatly increased our knowledge of gene and genome sequences. These data have accelerated greatly the pace of discovery in the biological sciences. However, even more rapid progress could be achieved if the interpretation of gene function from genome sequences were improved. Here we report a sequence search tool, CoSMoS, that is designed to perform context sensitive motif searches. Specifically, this algorithm is designed to recognize contiguous or gapped sequence motifs at constrained distances or ranges relative to secondary sequence-specified features. This search tool should be particularly useful for predicting patterns of protein distribution and gene expression because localization signals and transcription factor binding sites are specified by context-dependent sequence motifs.

As a first test of CoSMoS we searched for peroxisome-associated proteins of *Saccharomyces cerevisiae.* Peroxisomes are single membrane-bound organelles with a complex array of matrix proteins that are involved in a variety of metabolic processes (1, 2). Almost all peroxisomal matrix proteins carry the type 1 or type 2 peroxisomal targeting signals (PTS1 and PTS2, respectively) (3). These signals conform to relatively simple sequence motifs that function in particular sequence contexts. The PTS1 is an obligate C-terminal tripeptide of the sequence Ser-Lys-Leu-$_{COOH}$ (or a conservative variant), whereas the PTS2 has the consensus RL(X_5)HL and is located near the N terminus of proteins. In addition, peroxisomes are the sole site of fatty acid β-oxidation in yeast, and fatty acids regulate the expression of most *S. cerevisiae* peroxisomal proteins (4, 5). Fatty acid-dependent control of gene expression is mediated by the transcription factors PIP2 and OAF1 and transmitted via oleate response elements (OREs) in the promoters of responsive genes (6–8). The sequence CGG(N_3)TNA($N_{7–13}$)(G/C)CG represents a compromise between different consensus sequences that have been proposed for the ORE and is present in genes that encode many peroxisomal enzymes and biogenesis factors. We report here the use of CoSMoS, its application for the identification of peroxisomal proteins and oleate-regulated genes, and the discovery of at least 10 peroxisomal proteins, as well as a metabolic role for peroxisomes in lysine biosynthesis.

METHODS

Computer Analysis. We used PERL programming language to develop software systems capable of scanning a sequence database at the DNA or protein level for a given sequence in a context constrained manner. These programs are capable of scanning a protein or nucleotide database with a specific or degenerate sequence pattern at a designated position or range from or within an ORF or other sequence-derived element. The search programs also accommodate gaps of variable lengths within the sequence pattern searched. This search tool was used to survey a conceptually translated protein database of the entire *S. cerevisiae* genome sequence for the peptides (S/A/C)(K/H/R)L, (S/A)(Q/N)L, and SKF at the C terminus of proteins 100 amino acids or longer. CoSMoS was also used to scan a conceptually translated protein database of the *S. cerevisiae* genome for the PTS2 consensus RL(X_5)HL within the first 25 amino acids of proteins 100 amino acids or longer. In addition, we used CoSMoS to scan the entire *S. cerevisiae*

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

PNAS is available online at www.pnas.org.

This paper was submitted directly (Track II) to the *Proceedings* office.
Abbreviations: CoSMoS, context sensitive motif searches; PTS, peroxisomal targeting signal; ORE, oleate response elements; GFP, green fluorescent protein; PMP, peroxisomal membrane protein.
‡Present address: Rosetta Inpharmatics, Kirkland WA 98034.
¶Present address: Center for Molecular Medicine, University of British Columbia, Vancouver V5Z 4H4, Canada.
‖To whom reprint requests should be addressed at: Department of Biological Chemistry, Johns Hopkins University School of Medicine, 725 North Wolfe Street, Baltimore, MD 21205. e-mail: stephen.gould@qmail.bs.jhu.edu.

genome for the ORE consensus sequence CGG(N_{6-12})T(N)A(N_3)(C/G)CG constrained to 500 bp upstream of ORFs 100 codons or longer. [CoSMoS can be obtained by contacting D.B. at XREF (http://www.ncbi.nlm.nih.gov/XREFdb/)].

Yeast Strains and Plasmids. Yeast strains were grown under standard conditions, and transformations were performed by using standard protocols (9). The localization of candidate peroxisomal proteins was performed in the *S. cerevisiae* strain FY86 (10) and a *pex3* derivative in which the *PEX3* ORF was replaced by the *HIS3* gene. The *S. cerevisiae* strain BY4733 (11) was used for disruption of candidate oleate-regulated genes and genes encoding candidate peroxisomal proteins. Mutants were generated by one-step PCR-mediated gene disruption (12). The S65T mutant of the green fluorescent protein (GFP) gene was obtained from R. Tsien (13). Oleate-inducible, URA3-based GFP fusion vectors pGFP-X and pX-GFP were used for the expression of GFP fusion proteins in yeast (M.T.G. and S.J.G., unpublished observations). PTS1 candidates were fused to the C terminus of GFP in pGFP-X. PTS2 and membrane protein candidates were fused at the N terminus of GFP in pX-GFP.

Localization of GFP Fusion Proteins. The subcellular distribution of GFP fusion proteins was determined by confocal fluorescence and phase contrast microscopy. Plasmids designed to express each ORF in fusion to GFP were used to transform FY86 and its *pex3* derivative to uracil prototrophy. Resulting strains were grown for 24 hr in minimal medium lacking uracil, washed once in minimal medium, and then incubated for 16–24 hr in rich medium supplemented with oleic acid and Tween 40 (1% yeast extract/2% bacto-peptone/0.2% oleic acid/0.02% Tween 40). Cells were harvested, attached to cover glasses coated with poly-L-lysine, and observed by conventional and confocal fluorescence microscopy. Images were captured by using confocal fluorescence microscopy.

Northern Blot Analysis. The strain BY4733 was maintained at mid-log phase growth for 24–48 hr in minimal oleate medium or minimal ethanol medium. Two-liter cultures of cells were then harvested at an OD_{600} of 1.0, and RNA was extracted by using standard procedures (9). Poly(A)$^+$ RNA was purified by using Dynabeads according to the manufacturer's directions (Dynal, Great Neck, NY). A total of 0.5 μg of poly(A)$^+$ RNA was loaded per lane, separated by denaturing agarose gel electrophoresis, and transferred to nylon membranes. Filters were prepared for hybridization and hybridized with radiolabeled DNA fragments by using standard protocols (9).

RESULTS

Identification of PTS-Containing Proteins. We directed CoSMoS to scan the *S. cerevisiae* genome for genes encoding PTS1-containing proteins. Specifically, we searched for all ORFs $\geq$ 100 codons in length which encoded proteins with the sequences (S/A/C)(K/H/R)L, (S/A)(Q/N)L, or SKF at their C terminus. Thirty-five candidates were identified. These included seven of the eight known PTS1-containing proteins in *S. cerevisiae*, the sole exception being FAA2 (14). Computer analysis of the 28 remaining candidate PTS1 proteins led us to discard 10 as probable false positives because they had either been localized to other cellular compartments or were highly similar to known nuclear, mitochondrial, or ribosomal proteins. Of the 18 remaining candidates (Table 1; YIR027C to YBR012C), 5 had been assigned gene names but none had been shown to encode peroxisomal proteins.

Previous experiments have established that GFP can be targeted to peroxisomes by either a PTS1 or a PTS2, or by fusion to a peroxisomal membrane protein (PMP) (15). To determine whether the PTS1 candidates identified above represented an enriched subset of peroxisome-associated proteins, several were tagged at their N terminus with GFP and expressed in both wild-type yeast (FY86) and a *pex3* derivative [*PEX3* is required for peroxisomal matrix protein import (16)]. Control experiments confirmed that GFP is a cytoplasmic protein, and that fusions between GFP and known peroxisomal proteins, CIT2 (17), PCS60 (18), and SPS19 (19) were targeted to peroxisomes in a *PEX3*-dependent manner (data not shown). Seven of the 14 candidate peroxisomal proteins were also targeted to peroxisomes as determined by punctate localization in FY86 and cytoplasmic distribution in the *pex3* strain (Fig. 1; Table 1). In addition, the fusion GFP–YGR077C also appeared peroxisomal in FY86 cells but was also distributed to punctate structures in the *pex3* strain. Of the six GFP fusions that did not localize to peroxisomes, four were cytosolic, one was nuclear, and one fusion (GFP–YBR204C) was associated with triglyceride droplets.

Table 1. List of selected PTS and ORE candidate genes

Gene	PTS1	ORE	Distribution	*pex*	*onu*	*lys*
YCR005C/CIT2	SKL	Yes	Peroxisomal*	No	No	No
YKR009C/FOX2	SKL	Yes	Peroxisomal	ND	ND	ND
YDR256C/CTA1	SKF	Yes	Peroxisomal	ND	ND	ND
YDL078C/MDH3	SKL	Yes	Peroxisomal	ND	ND	ND
YML024W/CAT1	AKL	Yes	Peroxisomal	ND	ND	ND
YBR222C/PCS60	SKL	Yes	Peroxisomal*	ND	ND	ND
YNL202W/SPS19	SKL	Yes	Peroxisomal*	No	P	No
YIR027C/AAT2	AKL	No	Peroxisomal*	No	P	No
YNL117W/MLS1	SKL	No	**Peroxisomal***	No	P	No
YIR031C/MLS2	SKL	No	**Cytoplasmic***	No	No	No
YIR034C/LYS1	SRL	No	**Peroxisomal***	No	No	Yes
YDR234W/LYS4	SQL	No	**Peroxisomal***	No	No	Yes
YGR077C	SKL	No	**Peroxisomal***	Yes	Yes	P
YGL184C	SKL	NO	**Peroxisomal***	No	No	No
YNL009W	SKL	Yes	Peroxisomal*	No	P	No
YGR154C	SKL	No	**Peroxisomal***	No	No	No
YBR204C	SKL	No	**Lipid Droplet***	ND	ND	ND
YLR109W	AHL	No	**Cytoplasmic***	ND	ND	ND
YEL029C	ARL	No	**Cytoplasmic***	ND	ND	ND
YIL094C	SRL	No	**Cytoplasmic***	ND	ND	ND
YDR449C	SKL	No	**Nuclear***	ND	ND	ND
YGL067W	SHL	No	ND	ND	ND	ND
YMR259C	SQL	No	ND	ND	ND	ND
YBL071C	SHL	No	ND	ND	ND	ND
YBR012C	SRL	No	ND	ND	ND	ND
YJR019C	No	Yes	**Peroxisomal***	No	P	No
YLR284C	No	Yes	**Peroxisomal***	No	P	No
YOR180C	No	Yes	**Peroxisomal***	No	P	No
YPR128C	No	Yes	**Peroxisomal***	No	No	No
YOL044W	No	Yes	Peroxisomal*	Yes	Yes	P

PTS1 refers to the form or absence of PTS1 in the protein. ORE refers to whether the corresponding gene contains a consensus ORE sequence. Proteins that were localized in this study are marked by an asterisk, and those proteins that had not previously been localized are in bold. Mutants that display defects in peroxisome biogenesis were designated *pex*, mutants defective in growth on oleic acid were designated *onu*, and lysine auxotrophs were designated *lys*. ND, not determined; no means it is not a mutant; yes means it is a mutant; and P means that the defect appears to be partial.

We also scanned the *S. cerevisiae* genome for genes capable of encoding proteins carrying the PTS2 consensus sequence, RL(X_5)HL, near their N terminus. Four ORFs were identified, two of which lay within larger ORFs and were not studied further. The two other PTS2-containing proteins were peroxisomal thiolase (YIL169C/POT1), the only known PTS2 protein of *S. cerevisiae*, and YDL022W, which encodes a glycerol-3-phosphate dehydrogenase, GPD1. The POT1–GFP fusion was targeted to peroxisomes in a *PEX3*-dependent manner, whereas the GPD1–GFP fusion was cytosolic in both wild-type and *pex3* cells (data not shown).

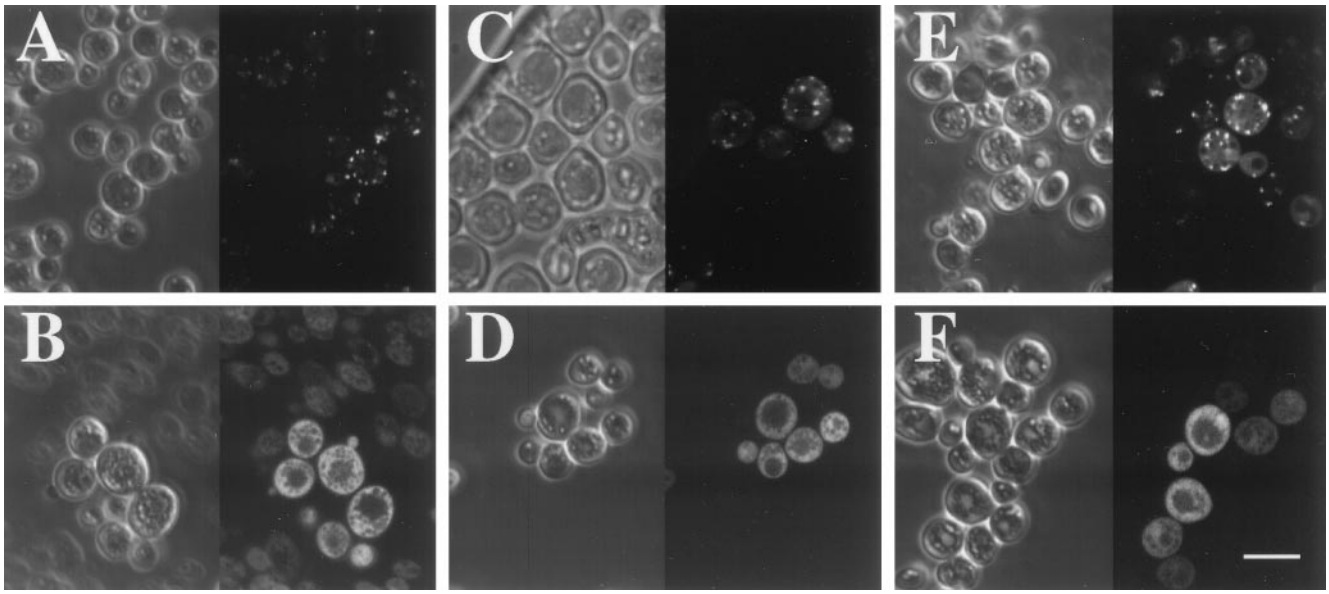

FIG. 1. Subcellular distribution of selected PTS1 candidate proteins. Confocal phase contrast and fluorescence microscopy of: (*A*) a wild-type and (*B*) a *pex3* strain expressing the GFP–YIR034C (*LYS1*) fusion protein; (*C*) a wild-type and (*D*) a *pex3* strain expressing the GFP- YDR234W (*LYS4*) fusion protein; (*E*) a wild-type and (*F*) a *pex3* strain expressing the GFP–YGL184C fusion protein. The left side of each panel shows the phase contrast image of the yeast cells and the corresponding fluorescence image is shown on the right side of each panel. (Bar = 5 μm.)

Identification of Oleate-Regulated Genes. We next examined the ability of CoSMoS to identify genes with common patterns of expression. Many peroxisomal proteins are encoded by oleate-regulated genes, most of which contain an ORE of the sequence CGG(N_3)TNA(N_{7-13})(G/C)CG. CoSMoS was used to identify all matches to this sequence in the *S. cerevisiae* genome that lay within 500 base pairs 5′ of ORFs ≥100 codons in length. A total of 340 OREs were identified in these putative promoters, with 290 genes carrying a single ORE, 22 genes carrying two OREs, and 2 genes carrying three OREs. Of the 314 genes identified in this scan, 122 had been assigned gene names previously, providing at least some concept of function. The other 192 genes are currently uncharacterized. The large number of genes identified in this scan makes their presentation here impractical but the data can be obtained from the authors (S.J.G. or M.T.G.).

CoSMoS correctly identified 13 of the 14 previously described oleate-regulated genes. Northern blots were used to determine whether any of the ORE candidates were also regulated by oleic acid. Specifically, mRNA levels in oleate-grown cells were compared with levels in ethanol-grown cells (Fig. 2). Five different candidate ORE-containing genes were examined (YLR284C, YPR128C, YOL044W, YOR180C, and YJR019C), as well as one non-ORE-containing gene (YEL020C). Four of the five candidate ORE-containing genes were more abundant in oleate-grown cells as compared with ethanol-grown cells (Fig. 2). Thus, a significant proportion of the candidate ORE-containing genes appear to be regulated by oleic acid.

In addition to Northern blot screens, we also initiated a computer-based screen of the ORE candidates for those which contained features expected of peroxisomal proteins. Specifically, we searched for proteins that (*i*) contained PTS1- or PTS2-like sequences or (*ii*) contained membrane spanning domains but lacked signal peptides and/or signal peptidase sites. Although this analysis is far from complete, we have identified three previously uncharacterized proteins that end in PTS1-like sequences: YOR180C, which terminates in HKL-$_{COOH}$, YLR284C, which terminates in HRL-$_{COOH}$, and YJR019C, which terminates in AKF-$_{COOH}$. These proteins were tagged at their N terminus with GFP, expressed in wild-type and *pex3* cells, and found to target GFP into peroxisomes (data not shown). The search for candidate PMPs was also successful, leading to the identification of YPR128C, an uncharacterized member of the ATP/ADP carrier family of solute transporters, and YOL044W, an orphan protein. Both proteins targeted GFP to peroxisomes in both wild-type and *pex3* cells, results that are consistent with a peroxisomal membrane localization for both proteins (data not shown).

Functional Studies of Novel Peroxisomal Proteins. To examine possible functions of these peroxisomal proteins, the corresponding genes were disrupted and the resulting strains were examined for defects in peroxisome biogenesis and growth on fatty acids, the two most prominent peroxisomal functions. GFP fusions that are targeted to peroxisomes via the PTS1, PTS2, and PMP targeting pathways (15) were expressed in all disruption mutants, and the biogenesis of peroxisomes was assessed by fluorescence microscopy. The only strains that

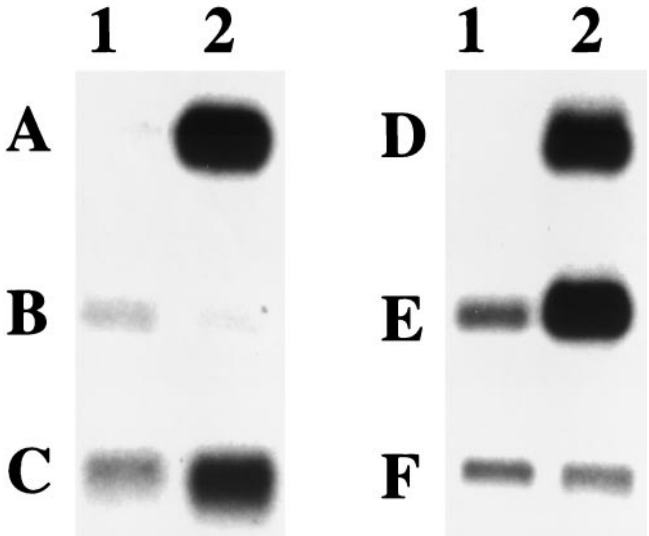

FIG. 2. Northern blot analysis of selected ORE-containing candidate genes. Poly(A)$^+$ RNA (0.5 μg) from ethanol-grown cells (lane 1) and oleate-grown cells (lane 2) were analyzed by Northern blot by using DNA probes specific for the genes (*A*) YLR284C, (*B*) YPR128C, (*C*) YOL044W, (*D*) YOR180C, (*E*) YJR019C, and (*F*) YEL020C.

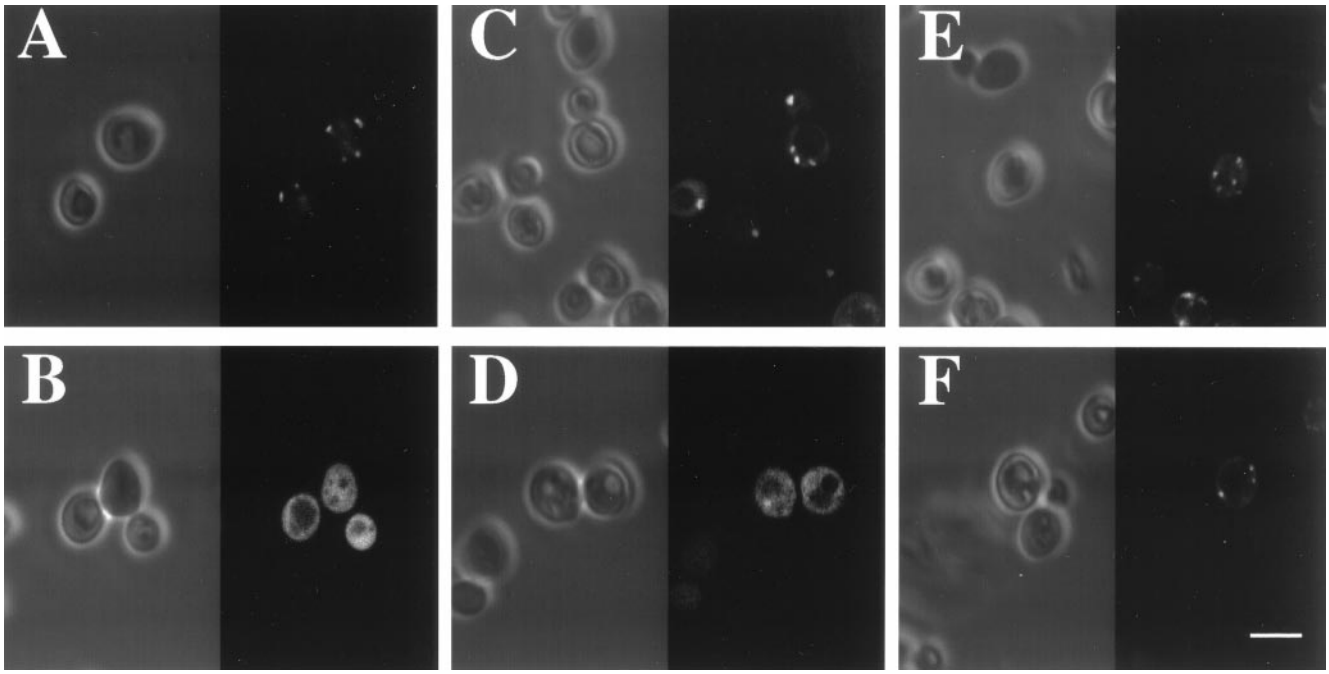

FIG. 3. YGR077C is required for peroxisomal matrix protein import. The wild-type strain BY4733 (*A*, *C*, and *E*) and a *HIS3* derivative of this strain lacking the YGR077C ORF (*B*, *D*, and *F*) were each modified so as to express PTS1-GFP (*A* and *B*), PTS2–GFP (*C* and *D*), and PMP–GFP (*E* and *F*). Each strain was then examined by confocal phase contrast (*Left*) and fluorescence (*Right*) microscopy. (Bar = 5 μm.)

displayed a defect in peroxisome biogenesis were those which lacked YGR077C (Fig. 3) or YOL044W (data not shown). These strains failed to import peroxisomal matrix proteins but showed no discernible defect in import of PMPs or the proliferation of peroxisomes in response to oleic acid. Thus, YGR077C and YOL044W appear to play roles in peroxisomal matrix protein import. The YGR077C gene product is most similar to PEX8 proteins identified in other yeasts [*Pichia pastoris* (20), *Hansenula polymorpha* (21), and *Yarrowia lipolytica* (22)], and we therefore designate YGR077C as the *S. cerevisiae PEX8* gene. Both YGR077C and YOL044W behaved as PMPs, a localization that is consistent with their roles in peroxisome biogenesis. Elgersma *et al.* (23) also identified YOL044W as a PMP required for peroxisome biogenesis, a result which led to its recent designation as *PEX15*.

In yeast, all fatty acid β-oxidation occurs in peroxisomes (4, 5). To examine possible roles for each gene in the metabolism of fatty acids we measured the growth rate of each strain in medium containing oleic acid (Δ9,C18:1) as sole carbon source (Fig. 4; Table 1). As expected for *S. cerevisiae pex* mutants, the *pex8* and the *pex15* strains failed to grow on oleic acid. The mutant lacking YLR284C also failed to grow on oleic acid. The *yir027c* (*aat2*), *ynl117w* (*mls1*), *ynl009w*, *yor180c*, and *yjr019c* mutants displayed intermediate phenotypes on oleic acid, indicating that the corresponding proteins may play ancillary roles in fatty acid utilization. Of these, the *aat2* mutant displayed the most interesting phenotype. Although this mutant was unable to grow at early time points, it began to grow on day 5 and its final density approached 50% of the wild-type strain.

LYS1 and LYS4 were the most unexpected of the peroxisomal proteins that were identified in our screen. These enzymes are responsible for the fourth and 11th (last) steps of lysine synthesis, respectively. Their peroxisomal localization suggested that lysine synthesis may represent a novel peroxisomal metabolic pathway. Therefore, all available disruption mutants were also tested for growth in the absence of exogenous lysine. In addition to the *lys1* and *lys4* mutants, which were lysine auxotrophs, the *pex8* and *pex15* mutants showed impaired growth on lysine-deficient solid medium, providing some additional evidence for a peroxisomal contribution to lysine synthesis.

DISCUSSION

Gene and genome sequences specify the function, expression, and distribution of most gene products. The elucidation of complete genome sequences raises the possibility of *in silico* deduction of gene functions on a genome-wide basis. However, the tools necessary to extract the biologically relevant information that is present in DNA and protein sequences have yet to be developed. Much of the information in sequence data exists in context-dependent forms, and we have developed CoSMoS to integrate context sensitivity constraints with current motif scanning strategies. Our results indicate that CoSMoS is an effective tool for gene identification. Furthermore, these studies have broadened our understanding of peroxisome content and function.

One approach to judging CoSMoS is to evaluate its accuracy. Specifically, what were the false-negative and false-positive identification rates? CoSMoS identified all known proteins that contain the forms of the PTS1 or PTS2 that were specified in the search. This included all but one of the eight known PTS1 proteins and the only known PTS2 protein of *S. cerevisiae*. In addition, CoSMoS identified 13 of the 14 known ORE-containing genes in this yeast. Thus, the searches appeared to have an acceptable false-negative rate. Although our current data are too limited to allow a precise measurement of the false-positive identification rates it is possible to extrapolate from the existing data set. Seventeen of the 37 proteins identified in the PTS1 and PTS2 scans were either known to be peroxisomal or shown to be peroxisomal in this report, and thus the upper limit to the false-positive rate of these scans is 54%. Northern blot analysis suggests an even lower false-positive rate for the ORE scan, in which four of the five genes tested were shown to be induced by oleic acid. Given that a 50% false-positive rate is quite acceptable for traditional genetic screens and selections, it would appear that CoSMoS strikes a balance of low false-negative rate with an acceptable false-positive rate. It should be noted that a similar computer-

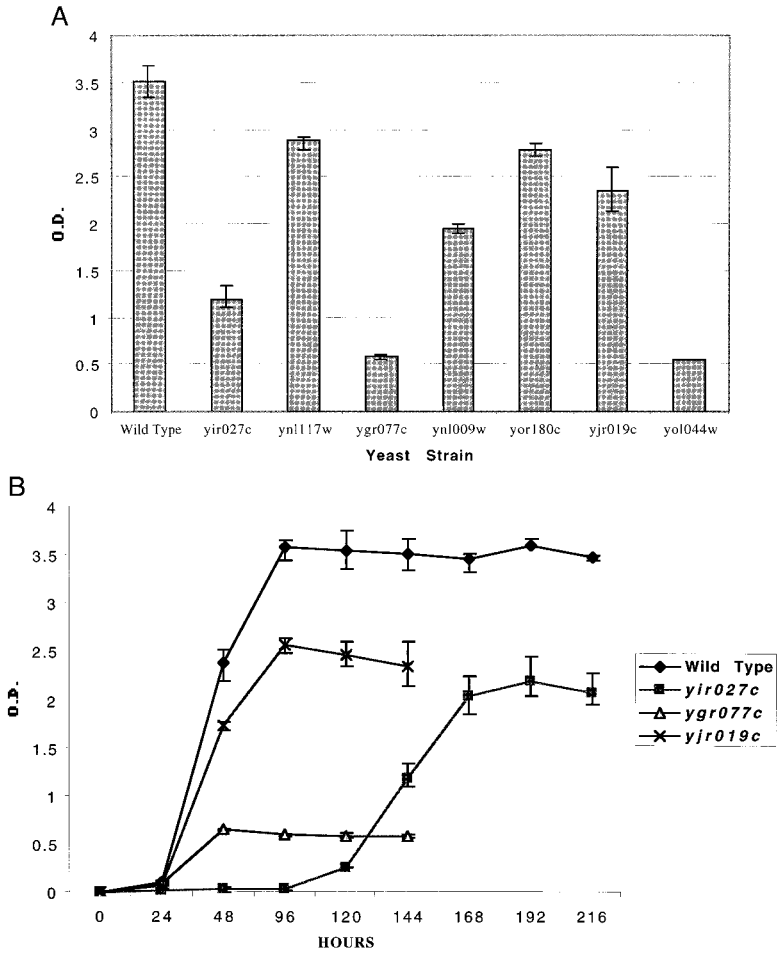

FIG. 4. Ability of different disruption mutants to grow on oleic acid as sole carbon source. The wild-type strain BY4733 and a series of *HIS3* derivatives lacking different PTS and ORE candidate genes were examined for their ability to grow in oleic acid-containing medium. Cells from each strain were maintained in mid-log phase growth in glucose-containing medium for 1 day and then used to inoculate a liquid oleic acid-containing culture at an OD_{600} of 0.01. Growth was monitored by following the change in absorbance at OD_{600}. Final growth densities for each strain (*A*) are presented as the average and absolute range of three trials. Detailed growth curves (*B*) for wild-type and a representative set of mutants are also presented as the average culture density ± the absolute range at each time point. Note the odd shape of the growth curve obtained for the *AAT2* mutant, *yir027C*, which grows on oleic acid only after a long lag period.

based search for oleate-regulated genes (24) has identified some of the genes reported here.

Although CoSMoS identified almost all known peroxisomal proteins and oleate-regulated genes, it is clear that the search parameters described in this study can be improved. Several nonperoxisomal proteins contain PTS sequences and several peroxisomal matrix proteins were excluded from the search. We are currently examining peroxisomal and nonperoxisomal PTS-containing proteins to identify other variables with predictive value, such as the relative hydrophilicity or charge of sequences upstream of the PTS1 and downstream of the PTS2. Likewise, we are examining the sequence context of OREs to determine whether other sequence features affect the functionality of these elements. Incorporation of refined search parameters in future screens may allow for the identification of even more candidate peroxisome-associated proteins and oleate-regulated genes.

In addition to demonstrating the utility of CoSMoS as a tool for gene identification, this study has significantly expanded our understanding of peroxisome biology. Thirteen peroxisomal proteins were identified in this study, although three of these [YIR027C/AAT2 (25), YNL009W/IDP3 (26, 27), and YOL044W/PEX15 (23)] have been reported recently by other groups while this paper was in preparation. These 13 proteins have roles in an array of processes, ranging from the oxidation of fatty acids and the synthesis of lysine to the biogenesis of peroxisomes. Of all of the candidates, LYS1 and LYS4 were perhaps the most unexpected because they define a metabolic role for the peroxisome in lysine synthesis. YGL184C, YGR154C, and YPR128C are also peroxisomal proteins but are of unknown function. In addition to peroxisomal constituents, CoSMoS identified candidates for enzymatic activities that were known previously or hypothesized to reside in peroxisomes. These include an NADP-dependent isocitrate dehydrogenase (YNL009W/IDP3), an aspartate aminotransferase (YIR027C/AAT2), a putative thioesterase (YJR019C), two enoyl-CoA hydratase-like proteins (YLR284C and YOR180C), and a putative malate synthase (YNL117W/MLS1).

Loss of any of YJR019C, YOR180C, YLR284C, and YNL117W(MLS1) resulted in impaired growth on fatty acids. YJR019C is highly similar to the human HIV Nef-associated thioesterase (28, 29). Current models do not incorporate thioesterase activity as an important feature of peroxisomal β-oxidation, and thus the phenotype of the *yjr019c* mutant is perplexing. However, it may be that efficient peroxisomal β-oxidation requires the ability to closely regulate the intraperoxisomal CoA levels. In light of this possibility it is interesting that loss of the opposing enzymatic activity, peroxisomal acyl-CoA synthetase (FAA2), results in a growth defect on fatty acids that is similar to that of the *yjr019c* mutant. YOR180C and YLR284C resemble proteins in the hydratase/

isomerase gene family (30) and we have found that YLR284C encodes a 2,3-enoyl-CoA isomerase (31). As for YNL117W-(MLS1), this protein is highly similar to malate synthases, enzymes which perform the second unique step of the glyoxylate cycle. Although the glyoxylate cycle is required for growth on fatty acids and traditionally thought of as a peroxisomal process, recent studies have raised doubt about these assumptions (32). MLS1 was clearly peroxisomal, but its loss resulted in only a mild growth defect on fatty acids. These data suggest that the role(s) of MLS1 and peroxisomes in the glyoxylate cycle require further investigation. In addition, we report here that YGR077C is closely related to PEX8 proteins of other species. It is interesting to note that PEX8 behaves as a PMP even though it contains a PTS1 signal (SKL_{COOH}) at its C terminus.

The ability of CoSMoS to detect virtually all known peroxisomal proteins and many peroxisomal proteins of *S. cerevisiae* demonstrates that CoSMoS can significantly impact our understanding of biological processes. In addition to being used for the identification of peroxisome-associated proteins it is also possible that CoSMoS could aid in identifying candidate endoplasmic reticulum proteins (containing the endoplasmic reticulum retention signals KDEL or HDEL), farnesylated or geranyl-geranylated proteins (containing a CAAX box at their C terminus), secreted proteins (containing hydrophobic stretch followed by a signal peptidase site), or mitochondrial proteins (basic, amphipathic stretch near the N terminus). Formulation of appropriate detection criteria for these motifs would expand our ability to deduce patterns of protein distribution from sequence data alone. Likewise, CoSMoS should be applicable to the identification of other coordinately regulated genes that share specific *cis*-acting DNA elements. Continued development and application of CoSMoS and related search tools should improve our ability to extract biologically relevant information from genetic sequences, a crucial step in maximizing the impact of genome sequence data.

This study was supported by a grant from the National Institutes of Health (DK45787) to S.J.G. and institutional funds from the Department of Pediatrics, the Johns Hopkins University School of Medicine to M.T.G.

1. Lazarow, P. B. & Fujiki, Y. (1985) *Annu. Rev. Cell Biol.* **1,** 489–530.
2. van den Bosch, H., Schutgens, R. B. H., Wanders, R. J. A. & Tager, J. M. (1992) *Annu. Rev. Biochem.* **61,** 157–197.
3. Subramani, S. (1993) *Annu. Rev. Cell Biol.* **9,** 445–478.
4. Kunau, W.-H. (1992) in *New Developments in Fatty Acid Oxidation* (Wiley–Liss, New York), pp. 9–18.
5. Kunau, W.-H., Beyer, A., Franken, T., Gotte, K., Marzioch, M., Saidowsky, J., Skaletz-Rorowski, A. & Wiebel, F. F. (1993) *Biochimie* **75,** 209–224.
6. Einerhand, A. W. C., Kos, W. T., Distel, B. & Tabak, H. F. (1993) *Eur. J. Biochem.* **214,** 323–331.
7. Rottensteiner, H., Kal, A. J., Filipits, M., Binder, M., Hamilton, B., Tabak, H. F. & Ruis, H. (1996) *EMBO J.* **15,** 2924–2934.
8. Karpichev, I. V., Luo, Y., Marians, R. C. & Small, G. M. (1997) *Mol. Cell. Biol.* **17,** 69–80.
9. Sambrook, J., Fritsch, E. & Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Lab. Press, Plainview, NY) 2nd Ed.
10. Winston, F., Dollard, C. & Ricupero-Hovasse, S. L. (1995) *Yeast* **11,** 53–55.
11. Baker-Brachmann, C., Davies, A., Cost, G. J., Caputo, E., Li, J., Hieter, P. & Boeke, J. D. (1998) *Yeast* **14,** 115–132.
12. Lorenz, M. C., Muir, R. S., Lim, E., McElver, J., Weber, S. C. & Heitman (1995) *Gene* **158,** 113–117.
13. Helm, R., Cubitt, A. B. & Tsien, R. Y. (1995) *Nature (London)* **373,** 663–664.
14. Elgersma, Y., Vos, A., van den Berg, M., van Roermund, C. W. T., van der Sluijs, P., Distel, B. & Tabak, H. F. (1996) *J. Biol. Chem.* **271,** 26375–26382.
15. Kalish, J. E., Keller, G. A., Morrell, J. C., Mihalik, S. J., Smith, B., Cregg, J. M. & Gould, S. J. (1996) *EMBO J.* **15,** 3275–3285.
16. Höhfeld, J., Veenhuis, M. & Kunau, W. H. (1991) *J. Cell Biol.* **114,** 1167–1178.
17. Lewin, A. S., Hines, V. & Small, G. M. (1990) *Mol. Cell. Biol.* **10,** 1399–1405.
18. Blobel, F. & Erdmann, R. (1996) *Eur. J. Biochem.* **240,** 468–470.
19. Gurvitz, A., Rottensteiner, H., Kilpelainen, S. H., Hartig, A., Hiltunen, J. K., Binder, M., Dawes, I. W. & Hamilton, B. (1997) *J. Biol. Chem.* **272,** 22140–22147.
20. Liu, H., Tan, X., Russell, K. A., Veenhuis, M. & Cregg, J. M. (1995) *J. Biol. Chem.* **270,** 10940–10951.
21. Waterham, H. R., Titorenko, V. I., Haima, P., Cregg, J. M., Harder, W. & Veenhuis, M. (1994) *J. Cell Biol.* **127,** 737–749.
22. Smith, J. J., Szilard, R. K., Marelli, M. & Rachubinski, R. A. (1997) *Mol. Cell. Biol.* **17,** 2511–2520.
23. Elgersma, Y., Kwast, L., van den Berg, M., Snyder, W. B., Distel, B., Subramani, S. & Tabak, H. F. (1997) *EMBO J.* **16,** 7326–7341.
24. Karpichev, I. V. & Small, G. M. (1998) *Mol. Cell. Biol.* **18,** 6560–6570.
25. Verleur, N., Elgersma, Y., Van Roermund, C. W., Tabak, H. F. & Wanders, R. J. (1997) *Eur. J. Biochem.* **247,** 972–980.
26. Henke, B., Girzalsky, W., Berteaux-Lecellier, V. & Erdmann, R. (1998) *J. Biol. Chem.* **273,** 3702–3711.
27. van Roermund, C. W., Hettema, E. H., Kal, A. J., van den Berg, M., Tabak, H. F. & Wanders, R. J. (1998) *EMBO J.* **17,** 677–687.
28. Watanabe, H., Shiratori, T., H, S., Miyatake, S., Okazaki, Y., Ikuta, K., Sato, T. & Saito, T. (1997) *Biochem. Biophys. Res. Commun.* **238,** 234–239.
29. Liu, L., Margottin, F., LeGall, S., Schwartz, O., Selig, L., Benarous, R. & Benichou, S. (1997) *J. Biol. Chem.* **272,** 13779–13785.
30. Muller-Newen, G., Janssen, U. & Stoffel, W. (1995) *Eur. J. Biochem.* **228,** 68–73.
31. Geisbrecht, B. V., Zhu, D., Schulz, K., Nau, K., Morrell, J. C., Geraghty, M., Schulz, H., Erdmann, R. & Gould, S. J. (1998) *J. Biol. Chem.* **273,** 33184–33191.
32. van Roermund, C. W., Elgersma, Y., Singh, N., Wanders, R. J. & Tabak, H. F. (1995) *EMBO J.* **14,** 3480–3486.